A Semantically-enabled Community Health Portal for Cancer Prevention and Control

Deborah L. McGuinness[1], Abdul R. Shaikh[2], Richard Moser[2], Bradford W. Hesse[2], Glen D. Morgan[2], Erik M. Augustson[2], Yvonne Hunt[2], Zaria Tatalovich[2], Gordon Willis[2], Kelly Blake[2], Paul Courtney[4], Lila Finney[4], Amy Sanders[4], Li Ding[1], Tim Lebo[1], Jim McCusker[1], Noshir Contractor[3], Yun Huang[3], York Yao[3], Hugh Devlin[3]

[1] Rensselaer Polytechnic Institute, Troy, NY

[2] National Cancer Institute, Rockville, MD

[3] Northwestern University, Evanston, IL

[4] SAIC-Frederick, Inc, NCI-Frederick, Frederick, MD

ABSTRACT

We describe our semantically-enabled approach to integrate, visualize, and explore health data. The project was conducted in a trans-disciplinary setting with population and behavioral scientists, social network scientists, data analysts, and computer scientists focused on making complex health-related data available, accessible, and understandable. One of the primary goals was to allow policy makers to explore potential correlations between health-related policies and behavior change. Other goals focused on demonstrating the value of linking open data and semantic technologies for exploration of data by research and consumer audiences. The initial setting includes comparison of smoking prevalence with potentially related data including cigarette taxes, price per pack, and policies limiting smoking in workplaces, restaurants, and bars, as well as personal information including education levels, employment, and various health statistics. The collaborative process, semantic data platform, demonstrations, and benefits of Linked Data for consumer data portals are also discussed.

Keywords

Semantic Web, Linked Data, Behavioral Science, Health Data Portal, Trans-disciplinary Collaboration

1. INTRODUCTION

In the face of a growing health care crisis, it is becoming more important to find ways to help researchers, policymakers, and consumers reduce disease and improve health. Our trans-disciplinary team is investigating ways to increase access to health-related data so that lay people and professionals can review, analyze, understand, and make informed decisions related to personal and population health issues. The team consists of members from a number of programs within the Division of Cancer Control and Population Science at the National Cancer Institute (NCI), associated contractors, and university researchers from Rensselaer Polytechnic Institute and Northwestern University with focus on the semantic technologies, social networks, and visualization. We collaboratively developed a Community Health Portal called PopSciGrid that explores the intersection of health behavior, policy, and demographic data using a Linked Data powered platform. While the initial domain revolves around tobacco policies, smoking prevalence, and related demographics, the platform can be used broadly in wide array of health data scenarios. Our work provides a foundation on which to base responses to calls for action such as those found in the President's Council of Advisors on Science and Technology's report to the President entitled "Realizing the full potential of health information technology to improve healthcare for Americans".[5]

2. BACKGROUND

Starting with a target area of tobacco-related health data exploration, the original team identified relevant datasets including data from the National Health Interview Survey (NHIS [1]) and the Health Information National Trends Survey (HINTS [2]), designed an initial portal [6] that integrated data from a portion of the data available, and began exploring potential connections. The team expanded in 2009 to include Semantic Technology and Linked Data experts who used their expertise from the Rensselaer Tetherless World Constellation Linked Open Government Data (LOGD) platform [2]. The new team together considered additional datasets from Impacteen.org (covering state-level statistics on smoking prevalence) and the National Cancer Institute (NCI) curated smoke-free policy coverage datasets. It also used Linked Data tools for conversion of data, integration, provenance, and visualization which resulted in an expanded and generalized portal [1] available on the web[3].

The project aims to provide data in an interactive, data-driven interface that enables people to easily explore factors that may impact smoking prevalence. For example, users may want to look for correlations between price per cigarette pack or tax per pack and smoking prevalence. Users may also want to view data associated with smoking bans in bars, restaurants, and workplaces and then they may wish to look for potential relationships

[1] http://www.cdc.gov/nchs/nhis.htm

[2] http://hints.cancer.gov/

[3] http://logd.tw.rpi.edu/project/popscigrid

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Semantically-enabled Community Health Portal For Cancer Prevention And Control	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Rensselaer Polytechnic Institute,Troy,NY,12180**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
To be presented at the Third International Conference on Web Science, Koblenz, Germany, June 14-17, 2011.,Government or Federal Purpose Rights License

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**4**	

between these bans and smoking prevalence. Figures 1 and 2 show two different interactive demonstrations available from our PopSciGrid portal that allow users to visually explore the integrated data. We use a variety of visualization tools, often using the Google visualization toolkit[4]. Figure 1 shows the result of running a motion chart over time displaying data indicating that tobacco tax per cigarette pack increased (on the y axis) and smoking bans increased over time (on the x axis). Note that smoking prevalence is indicated by the color and size of the circle representing individual states in the chart. In Figure 1, only the states that have been selected using the checkboxes on the bottom right are labeled in the graph. (This allows users to focus more easily on individual states and helps to reduce some visual clutter). When viewing this demonstration interactively, it becomes easy to see which states tried more aggressive policies and when they did so. One can also view associated smoking prevalence over time. The interactive version of this demonstration is available from http://logd.tw.rpi.edu/demo/tax-cost-policy-prevalence. Figure 2 shows a map of smoking prevalence by state. The graphs on the right are configurable to allow the user to select any state and view its smoking prevalence, smoke-free policy coverage, cigarette price, and tax over time. The interactive version of this demonstration is available directly through the LOGD web site.[5]

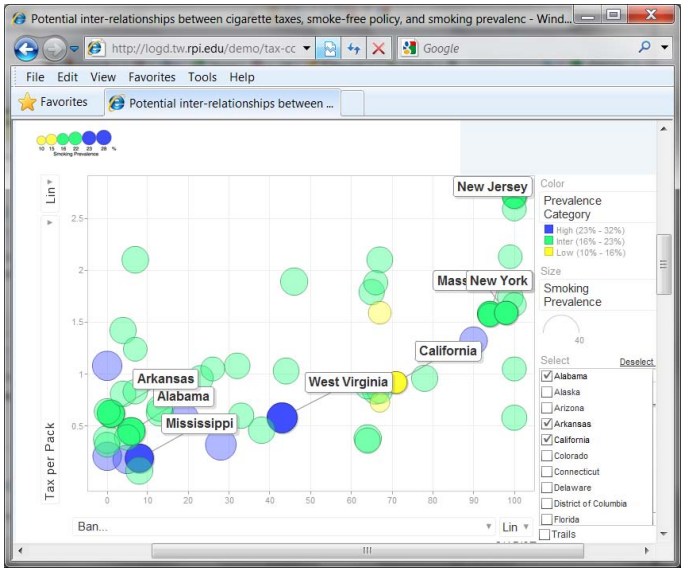

Figure 1. Taxation, Policy, and Prevalence over time.

3. Methods

Data were gathered from the ImpacTeen State Level Tobacco Control Policy and Prevalence Database [3] covering state tobacco control policy and prevalence data. Data were also gathered from the National Cancer Institute covering changes in tobacco ban and taxation policy based on the Chronological Table of U.S. Population Protected by 100% Smokefree State or Local Laws available from ANRF[6] from 1990 to 2007. A detailed description of the data gathering and analysis effort is in process in [7].

In order to build the portal, the Rensselaer researchers used automated conversion software[7] called CSV2RDF4LOD (Comma Separated Value to Resource Description Framework for Linked Open Data) to turn raw government data, formatted in CSV, into RDF triples that follow the Linked Data principles. This converter is similar to the RDF converter extension available for Google Refine[8], however the Google Refine RDF extension is designed for interactive mapping of RDF data, and CSV2RDF4LOD is optimized for use in large scale, repeatable conversions of data. These data are then stored in an RDF database or 'triple store'. We use OpenLink Virtuoso[9], but any SPARQL-enabled triple store is sufficient to the task. The RDF converter also tracks provenance: i.e., it maintains the association of the converted RDF data, with the initial raw data and further includes the conversion configuration. It encodes the provenance information using PML Provenance Interlingua [4]. Capturing the provenance along with the raw data allows applications to query for provenance information along and potentially support interfaces and applications that filter using provenance (such as only obtaining data from particular sources, or using particular kinds of analyses).

In the converted data, variables (such as smoking prevalence, cigarette taxation, and smoke-free policy coverage) and dimensional parameters (such as state and year) are identified. The integration of datasets is relatively straightforward. For example, we routinely join datasets by state (or geographic region) and by time periods (such as year or month). The created visualizations, written in JavaScript, query the triple store using SPARQL and retrieve results in JSON which are immediately consumable by Google Visualization APIs. The two featured demonstrations use Google visualizations, although other visualization tools have been used on these data as well.

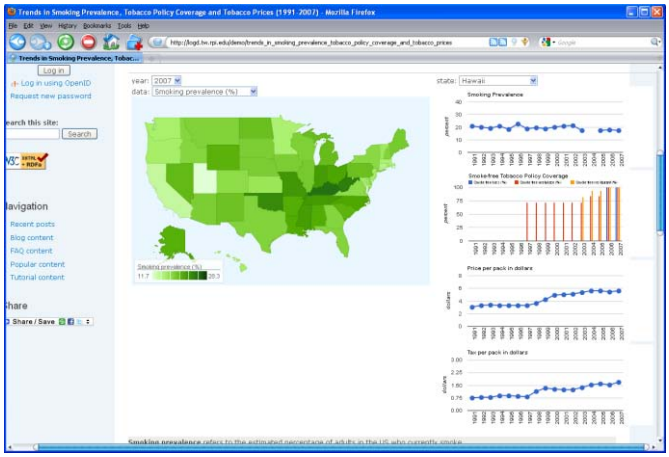

Figure 2. Smoking prevalence by state with graphs for prevalence, ban policies, price, and tax.

[4] http://code.google.com/apis/chart/interactive/docs/gallery.html

[5] http://logd.tw.rpi.edu/demo/trends_in_smoking_prevalence_tobacco_policy_coverage_and_tobacco_prices

[6] http://www.no-smoke.org/pdf/EffectivePopulationList.pdf

[7] http://logd.tw.rpi.edu/technology/csv2rdf4lod

[8] http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/

[9] http://virtuoso.openlinksw.com/

4. Discussion

The results of the so-called mashup of the data are visualized in the figures. We found it relatively simple to take the initial hand generated demonstration system and replicate it in a more automated, extensible, and scalable manner. Even though these demonstrations only display a small set of the parameters in our triple store, it is straightforward to allow users to explore relationships between any of the parameters in the triple store. For example, we have explored relationships between education level, job status (employed or unemployed), self-reported depression levels, and smoking prevalence. By using the automated tools, we have a consistent and relatively complete encoding of provenance that we can expose in our demonstrations and search interfaces. Furthermore, the variety of visualization tools that can be generated from the platform is significant, so we can support alternative views as needed. The ease of creating multiple visualizations allowed us to view data in multiple ways. One side effect of these visualizations was that it allowed for identification of some anomalies that became obvious when viewing data in specific forms (e.g., some visualizations clearly identified values above 100% which led us to investigate and discover some rounding errors). We were able to utilize the recorded provenance trace as the basis of communication with government data curators to investigate and fix issues in the data processing. We have found that this flexibility in demonstration format and content to be particularly useful in working in a broad trans-disciplinary team.

Our work on health information portals is in its early stages. These demonstrations are aimed at exploration and hypothesis formation for further investigation. We intend to expand our initial demonstration to include other data, including additional demographics and possibly relevant health statistics such as lung cancer data. We are also exploring additional visualizations. In addition, we plan to use the same platform to generate demonstrations in other key health-related behaviors. For example, the Rensselaer team has obtained the data from the NCI CLASS project – Classification of Laws with School Students[10] to begin to explore potential relationships between policies related to physical education requirements and nutrition in schools with the intention to create visualizations with the potential to explore weak and strong policy regions. These regions could then be mashed up against data related to childhood obesity and possible related health issues. Similarly, we have begun to collect data by state and sometimes by zip code, providing further granularity for use in mashing up against policy data, thus enabling lay people as well as policy makers to investigate possible trends in policies by regions and potential correlations in health data and demographics in the same regions.

5. ACKNOWLEDGEMENTS

The authors wish to acknowledge the following funding sources that contributed to the development of materials presented in this poster: Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053; and National Science Foundation Grants: CNS-1010904, OCI-0904356, IIS-0838564, and IIS-0836262. Additionally, this project has been funded in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. This work was also made possible in part by gifts from Lockheed Martin, Fujitsu Labs America, LGS, and Microsoft. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services nor other entities, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government nor other institutions.

[10] http://class.cancer.gov/About.aspx

6. REFERENCES

[1] P. Courtney, A. R. Shaikh, N. Contractor, D.L. McGuinness, L. Ding, E. M. Augustson, K. Blake, G. D. Morgan, R. Moser, G. Willis, B. W. Hesse, Consumer Health Portal: An Informatics Tool for Translation and Visualization of Complex, Evidence-Based Population Health Data for Cancer Prevention and Control, In 138th APHA Annual Meeting, (2010).

[2] L. Ding, D. DiFranzo, A. Graves, J. Michaelis, X. Li, D.L. McGuinness, and J. Hendler (2010). TWC data-gov corpus: incrementally generating linked government data from data.gov. In 19th Intl World Wide Web Conference.

[3] G. A. Giovino, F. J. Chaloupka, A. M. Hartman, K. Gerlach Joyce, J. Chriqui, C. T. Orleans, K. Wende, C. Tworek, D. Barker, J. T. Gibson, J. Yang, J. Hinkel, K. M. Cummings, A. Hyland, B. Fix, M. Paloma, M. Larkin. Cigarette Smoking Prevalence and Policies in the 50 States: An Era of Change. – The Robert Wood Johnson Foundation ImpacTeen Tobacco Chart Book. Buffalo, NY: University at Buffalo, State University of New York, 2009. http://impacteen.org/statetobaccodata/chartbook_final060409.pdf

[4] D. L. McGuinness, L. Ding, P. Pinheiro da Silva, and C. Chang. PML 2: A Modular Explanation Interlinga. In Proc. Of the AAAI '07 Workshop on Explanation Aware Computing, July 2007. ftp://ftp.ksl.stanford.edu/pub/KSL_Reports/KSL-07-07.pdf

[5] President's Council of Advisors on Science and Technology. Realizing the full potential of HIT to improve healthcare for Americans. Washington: Executive Office of the President, 2010. http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf

[6] A. R. Shaikh, N. Contractor, R. Moser, G. D. Morgan, P. K. Courtney, E. M. Augustson, A. M. Pilsner, B. W. Hesse. PopSciGrid: Using cyberinfrastructure to enable data harmonization, collaboration, and advanced computation of nationally representative behavioral, demographic, and economic data. In 137th APHA Annual Meeting, (2009).

[7] Z. Tatalovich, Y. Hunt, S. Marcus, N. Howlader, and A. Mariotto. Geographic Patterns of Local Tobacco Control Ordinances and Cigarette Consumption in the US. In process.